

Transition States in Protein Folding Kinetics: The Structural Interpretation of Φ -values

Thomas R. Weikl¹ and Ken A. Dill²

¹*Max Planck Institute of Colloids and Interfaces, Theory Department,
14424 Potsdam, Germany*

²*Department of Pharmaceutical Chemistry, University of California,
San Francisco, California 94143-2240, USA*

Abstract

Φ -values are experimental measures of the effects of mutations on the folding kinetics of a protein. A central question is which structural information Φ -values contain about the transition state of folding. Traditionally, a Φ -value is interpreted as the ‘nativeness’ of a mutated residue in the transition state. However, this interpretation is often problematic because it assumes a linear relation between the nativeness of the residue and its free-energy contribution. We present here a better structural interpretation of Φ -values for mutations within a given helix. Our interpretation is based on a simple physical model that distinguishes between secondary and tertiary free-energy contributions of helical residues. From a linear fit of our model to the experimental data, we obtain two structural parameters: the extent of helix formation in the transition state, and the nativeness of tertiary interactions in the transition state. We apply our model to all proteins with well-characterized helices for which more than 10 Φ -values are available: protein A, CI2, and protein L. The model captures nonclassical Φ -values < 0 or > 1 in these helices, and explains how different mutations at a given site can lead to different Φ -values.

Introduction

There has been much interest in understanding the rates of protein folding in terms of transition state structures. We focus here on two-state proteins, i.e. those proteins that fold with single-exponential kinetics. The folding kinetics of two-state proteins is often investigated by mutational analysis [1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,23,24]. The effect of a

given mutation on the protein’s folding kinetics is quantified by its Φ -value [25,26]

$$\Phi = \frac{RT \ln(k_{\text{wt}}/k_{\text{mut}})}{\Delta G_N} \quad (1)$$

Here, k_{wt} is the folding rate for the wildtype protein, k_{mut} is the folding rate for the mutant protein, and ΔG_N is the change of the protein stability induced by the mutation. The stability G_N of a protein is the free energy difference between the denatured state D and the native state N . In classical transition-state theory, the folding rate of a two-state protein is proportional to $\exp[-G_T/RT]$, where G_T is the free energy difference from the denatured state to the transition state.¹ In that notation, Φ -values have the form

$$\Phi = \frac{\Delta G_T}{\Delta G_N} \quad (2)$$

where each Δ in this expression represents the change due to the mutation.

By definition, Φ -values are energetic quantities, related to changes in the protein’s stability and folding rate. Do Φ -values also give information about the structures that the protein adopts when it is in a kinetic “bottleneck” or transition state [26,27,28,29,30]? In the traditional interpretation, Φ -values are taken to indicate the *degree of structure formation of the mutated residue in the transition-state ensemble T*. A Φ -value of 1 is interpreted to indicate that the residue is fully native-like structured in T, since the mutation shifts the free energy of the transition state T by the same amount as the free energy of the native state N. A Φ -value of 0 is interpreted to indicate that the residue is as unstructured in T as in the denatured state D, since the mutation does not shift the free energy difference between these two states. Φ -values between 0 and 1 are taken to indicate partial native-like structure in T.

Modelers often calculate Φ -values based on this traditional interpretation. In many approaches, Φ -values are calculated from the fraction of contacts a residue forms in the transition state T, compared to the fraction of contacts in the native and the denatured state [31,32,33,34,35,36,37,38,39,40,41,42,43,44], or from similar structural parameters [45,46]. Notable exceptions are a recent MD study of an ultrafast mini-protein in which Φ -values are calculated from rates for the wildtype and mutants via eq. (1) [47], and the calculation of Φ -values from free energy shifts of the transition-state ensemble using eq. (2) [48].

¹ In principle, the prefactor of this proportionality relation could also depend on the mutation, but this dependence is usually neglected.

However, there are reasons to question this simple interpretation of Φ -values. First, some Φ -values are negative or larger than 1 [49,50]. These ‘nonclassical’ Φ -values cannot be interpreted as a degree of structure formation, because this would have the nonsensical implication of ‘less structured than D ’ or ‘more structured than N ’. Second, Φ -values are sometimes significantly different for different mutations at a given chain position, contradicting the normal assumption that the degree of nativeness of the transition state is just a property of the position of a monomer in the protein. Third, Φ -values for neighboring residues within a given secondary structure often span a wide range of Φ -values. In the traditional interpretation, this means that some of the helical residues are unstructured in the transition state, while other residues, often direct neighbors, are highly structured. This contradicts the notion that secondary structures are cooperative.

The inconsistencies of the traditional interpretation result from the assumption that the mutation-induced free energy changes of a residue are proportional to a single structural parameter, the ‘degree of nativeness’ of this residue in the transition state T . Is there a consistent structural interpretation of Φ -values, and if yes, how many structural parameters do we need to capture the mutation-induced free energy changes? We show here that the Φ -values for multiple mutations in a given helix can be consistently interpreted in a simple physical model that takes into account just two structural parameters for the whole helix: χ_α , the *degree of secondary structure formation of the helix in the transition-state ensemble T* , and χ_t the *degree of tertiary structure formation of the helix in T* . In our model, the mutation-induced free energy changes are split into two components. The overall stability change ΔG_N is split into two parts: the change in intrinsic helix stability ΔG_α , and the change in tertiary free energy ΔG_t caused by the mutation. Similarly, ΔG_T , the change of the free energy difference between the transition state and the denatured state, is split into a change $\chi_\alpha \Delta G_\alpha$ in secondary free energy, and a change $\chi_t \Delta G_t$ in tertiary free energy. The Φ -values for the mutations in the helix then have the general form

$$\Phi = \frac{\chi_\alpha \Delta G_\alpha + \chi_t \Delta G_t}{\Delta G_N} = \chi_t + (\chi_\alpha - \chi_t) \frac{\Delta G_\alpha}{\Delta G_N} \quad (3)$$

The second expression simply results from replacing ΔG_t by $\Delta G_N - \Delta G_\alpha$. The two parameters χ_α and χ_t of our model are ‘collective’ structural parameters for all mutations in the helix. Different Φ -values then simply result from different free-energetic ‘signatures’ ΔG_α and ΔG_N of the mutations. In particular, eq. (3) captures that different mutations of the same residue can lead to different Φ -values, and that Φ -values can be ‘nonclassical’, i.e. < 0 or

> 1 . Since the two structural parameters χ_α and χ_t range between 0 and 1, a nonclassical Φ -value implies that the changes ΔG_α and ΔG_t in secondary and tertiary free energy caused by the mutation have opposite signs.

To apply our model, we first estimate ΔG_α , the change in helical stability, for each mutation in a particular helix, using standard helix propensity methods. We then plot all experimental values for Φ versus $\Delta G_\alpha/\Delta G_N$, and obtain the two structural parameters χ_α and χ_t from a linear fit of eq. (3). In principle, the two structural parameters can be extracted if Φ -values and stability changes for at least two mutations in a helix are available. However, to test our model, and to obtain reliable values for χ_α and χ_t , we focus here on helices for which more than 10 Φ -values have been determined. The modeling quality then can be assessed from the standard deviation of the data points from the regression line, and from the Pearson correlation coefficients between Φ and $\Delta G_\alpha/\Delta G_N$. Our model can be applied to all mutations for a helix, or to a subset of mutations that affect only the tertiary interactions with one other structural element.

Models and methods

Transition-state conformations and folding rate

We model the transition state as an ensemble of M different conformations (see Fig. 1). Each transition-state conformation is directly connected to the native state N and to the denatured state D. The model thus has M parallel folding and unfolding routes.

We assume that the protein is stable, i.e. that $G_N < 0$. We also assume that the free energy barrier for each transition state conformation is significantly larger than the thermal energy, i.e. that $G_m/RT \gg 1$ [51,52]. The rate of folding along each route m is then proportional to $\exp[-G_m/RT]$, and the total folding rate as the sum over all the parallel routes is

$$k = c \sum_{m=1}^M e^{-G_m/RT} \quad (4)$$

where c is a constant prefactor.²

² This model is a generalization of our previous model [53] with $M = 2$ transition-state conformations. The master equation that describes the folding kinetics of this

Decomposition of free energy changes for helical mutations

Consider all mutations $i = 1, 2, \dots$ within one particular α -helix of a protein. The effect of these mutations on the stability and folding kinetics can be experimentally characterized by the stability changes ΔG_N , and by the Φ -values. We suppose that the experimentally measured change in stability ΔG_N for each mutation is the sum of effects on the stability of the helix and on the interactions of the helix with its tertiary neighbors:

$$\Delta G_N = \Delta G_\alpha + \Delta G_t \quad (5)$$

The first term, ΔG_α , is the change in the intrinsic helix stability. The second term, ΔG_t , is the change in tertiary free energy of the helix interactions with neighboring structures. Below, we estimate ΔG_α using either the program AGADIR [54,55,56] or from a helix propensity scale [57]. The term ΔG_t is then simply obtained by subtracting ΔG_α from the experimentally measured stability change, ΔG_N .

We also decompose each ΔG_m , the mutation-induced free energy change for the transition state conformation m , into two terms:

$$\Delta G_m = s_m \Delta G_\alpha + t_m \Delta G_t \quad (6)$$

Here, s_m is either 0 or 1, depending on whether the helix is formed or not in the transition state conformation m . The coefficient t_m is between 0 and 1 and represents the degree of tertiary structure formation in conformation m .

Structural and energetic components of Φ -values

The folding rate for the mutant protein i is $k_{\text{mut}} = k(G_1 + \Delta G_1, G_2 + \Delta G_2, \dots, G_M + \Delta G_M)$ with k given in eq. (4). The folding rate of the wildtype is $k_{\text{wt}} = k(G_1, G_2, \dots, G_M)$. We assume here that the mutations do not affect the pre-factor c in eq. (4). For small values $|\Delta G_m|$ of the mutation-induced free-energy changes, a Taylor expansion of $\ln k_{\text{mut}}$ gives

$$\ln k_{\text{mut}} - \ln k_{\text{wt}} \simeq \sum_{m=1}^M \frac{\partial \ln k_{\text{wt}}}{\partial G_m} \Delta G_m = -\frac{1}{RT} \frac{\sum_m \Delta G_m e^{-G_m/RT}}{\sum_m e^{-G_m/RT}} \quad (7)$$

model can be solved exactly. Eq. (4) is obtained from the exact solution in the limit of large transition state barriers G_m [53].

With the decomposition of the ΔG_m 's in eq. (6), we obtain

$$\ln k_{\text{mut}} - \ln k_{\text{wt}} \simeq -\frac{1}{RT} (\chi_\alpha \Delta G_\alpha + \chi_t \Delta G_t) \quad (8)$$

with the two terms

$$\chi_\alpha \equiv \frac{\sum_m s_m e^{-G_m/RT}}{\sum_m e^{-G_m/RT}} \quad \text{and} \quad \chi_t \equiv \frac{\sum_m t_m e^{-G_m/RT}}{\sum_m e^{-G_m/RT}}. \quad (9)$$

The term χ_α represents the Boltzmann-weighted average of the secondary structure parameter s_m for the transition-state ensemble T. χ_α ranges from 0 to 1 and indicates the average degree of structure formation for the helix in T. The value $\chi_\alpha = 1$ indicates that the helix is formed in all transition-state conformations m , and $\chi_\alpha = 0$ indicates that the helix is formed in none of the transition-state conformations. Values of χ_α between 0 and 1 indicate that the helix is formed in some of the transition-state conformation, and not formed in others. The term χ_t represents the Boltzmann-weighted average of the tertiary structure parameter t_m in T, and also ranges from 0 to 1. From eq. (8) and the definition in eq. (1), we then obtain the general form (3) of the Φ -values for helical mutations in our model.³

More than twenty two-state proteins with α/β [1,2,3,4,5,6,7,8,9,10,11,12], α -helical [13,14,15,16], or all- β structures [17,18,19,20,21,22,23,24] have been investigated by mutational analysis in the past few years. Mutational data are also available for several proteins that fold via intermediates [58,59,60] or apparent intermediates [61]. We focus here on the well-characterized α -helices of two-state proteins for which at least 10 Φ -values apiece are available: the helices 2 and 3 from the protein A, and the helices of CI2 and protein L. Protein A is an α -helical protein with three helices, CI2 and protein L are α/β -proteins with a single α -helix packed against a β -sheet.

³ In principle, our parameter χ_t for the tertiary interactions can also be seen to depend on the residue position. To derive eq. (3), we don't have to assume that the tertiary parameters t_m for the m transition-state conformations are independent of the residue position and/or mutation. However, we focus here on the simplest version of our model and show that a consistent structural interpretation of experimental Φ -values in a helix can be obtained with just two structural parameters χ_α and χ_t for the whole helix, which implies a cooperativity of secondary as well as tertiary interactions.

Results and discussion

Our analysis of experimental Φ -values requires an estimate of the mutation-induced changes ΔG_α of the intrinsic helix stability. In the case of the CI2 helix, we estimate ΔG_α both with the program AGADIR [54,55,56] and from a helix propensity scale [57], see Table 1. The change in intrinsic helix stability ΔG_α can be estimated from the helical content predicted by AGADIR via $\Delta G_\alpha = RT \ln (P_\alpha^{\text{wt}}/P_\alpha^{\text{mut}})$. Here, P_α^{wt} is the helical content of the wildtype helix, and P_α^{mut} the helical content of the mutant. The program AGADIR is based on helix/coil transition theory, with parameters fitted to data from Circular Dichroism (CD) spectroscopy. In Table 1, the values for ΔG_α obtained from AGADIR are compared to values from a helix propensity scale [57]. Helix propensities of the amino acids are typically given as free energies differences with respect to Alanine. We use the propensity scale of Pace and Scholtz [57], which has been obtained from experimental data on 11 different helical systems. For example, the value $\Delta G_\alpha = 0.29$ kcal/mol for the mutant E15D in the CI2 helix is simply the difference between the helix propensity 0.69 kcal/mol for the amino acid D (Aspartic acid) and the propensity 0.40 kcal/mol for amino acid E (Glutamic acid). The helix propensity scale can be applied for residues at ‘inner’ positions’ of a helix, not for residues at the termini or ‘caps’ of the helix. The N-terminal residues of the CI2 helix are the residues 12 and 13, the C-terminal residues are the residues 23 and 24. For the 8 mutations at ‘inner positions’ of the CI2 helix, the values for ΔG_α from AGADIR and from the helix propensity scale correlate with a Pearson correlation coefficient of 0.77. For the other three helices considered here, the helicities predicted by AGADIR are significantly smaller than the helicities around 5 % predicted for the CI2 helix. Estimates for ΔG_α based on AGADIR therefore are not reliable for these helices. The values of ΔG_α shown in the Tables 2 to 4 are calculated from helix propensities.

The three structural elements of protein A are its three helices. Based on the contact map of protein A shown in Fig. 2, the mutations in helix 2 of protein A can be divided into three groups: ‘purely secondary’ mutations that don’t affect tertiary contacts; mutations that affect only tertiary contacts with helix 1; and mutations that affect tertiary contacts both with helix 1 and 3. If only the first two groups of mutations are considered in our analysis, χ_t represents the average degree of structure formation with helix 1. If all groups and, thus, all mutations are considered, χ_t is the average degree of structure formation with the helices 1 and 3. In the case of helix 3, we distinguish between mutations that affect either tertiary contacts with helix 1 or helix 2,

or none of the tertiary interactions, see Table 3. In the case of the protein L helix, the two other structural elements are the terminal β -hairpins, see Fig. 3 and Table 4. In the case of CI2, we do not distinguish between different tertiary contacts. One reason is that there are at least three other structural elements to consider, the three strand pairings $\beta_2\beta_3$, $\beta_3\beta_4$, and $\beta_1\beta_4$ of the four-stranded β -sheet that is packed against the CI2 helix [53]. Another reason is that the degree χ_t of tertiary structure formation in the transition state is small for this helix.

The structural parameters χ_α and χ_t obtained from our analysis shown in the Figs. 4 to 7 are summarized in Table 5. We estimate the overall errors of χ_α and χ_t , which result from experimental errors in Φ and ΔG_N and from modeling errors, as ± 0.05 for the CI2 helix and helix 2 of protein A, and as ± 0.1 for helix 3 of protein A and the protein L helix. The χ_α values for the CI2 helix and the helix 2 of protein A are close to 1. This indicates that the helices are fully formed in the transition-state ensemble. In contrast, χ_α for helix 3 of protein A is close to 0, indicating that the helix is not formed in the transition state. χ_α for the helix in protein L indicates a partial degree of helix formation between 20 and 30 %. Our χ_t values indicate that the degree of tertiary structure formation in the transition state is around 16 % for the CI2 helix, around 50 % for helix 2 of protein A, and around 30 % for helix 3 of protein A. The χ_t values for the protein L helix show a small degree of tertiary structure formation with hairpin 1 (around 15 %) and no tertiary structure formation with hairpin 2.

To assess the quality of our modeling, we consider two quantities: the correlation coefficient r , and the estimated standard deviation SD of the data points from the regression line. High correlation coefficients up to 0.9 and larger indicate a high quality of modeling. However, it's important to note that the correlation coefficient can only be used to assess the modeling quality in the cases where the structural parameters χ_α and χ_t are sufficiently different from each other. The case $\chi_\alpha = \chi_t$ corresponds to a regression line with slope 0, and hence a correlation coefficient of 0, irrespective of how well the data are represented by this line. For small differences of χ_α and χ_t , the correlation coefficient r is dominated by the experimental errors in Φ . This is the case for the mutations in the protein L helix that affect tertiary contacts with hairpin 1, see Table 5. The slope of the regression line is almost zero for this data set, see Fig. 7. Here, the relatively small standard deviation 0.1 of the data points from the regression lines indicates that our model is in good agreement with the experimental data.

We only consider here mutations with stability changes $\Delta G_N > 0.7$ kcal/mol. Because of experimental errors, Φ -values for mutations with smaller stability changes are generally considered as unreliable [63,24,64]. In our previous work [53], we considered all the published mutations for the CI2 helix, including those for which ΔG_N is significantly smaller than 0.7 kcal/mol. The correlation coefficient 0.91 obtained here for the subset of mutations with $\Delta G_N > 0.7$ kcal/mol is larger than the correlation coefficient 0.85 for all mutations. The significantly larger reliability threshold of 1.7 kcal/mol for ΔG_N obtained by Sanchez and Kiefhaber [62] is based on the assumption that different mutations at the same residue position should lead to the same Φ -value. In our model, different Φ -values for mutations at the same site result from different effects on the intrinsic helix stability G_α and the tertiary free energy $G_t = G_N - G_\alpha$.

In our model, nonclassical Φ -values < 0 or > 1 can arise if $\Delta G_\alpha/\Delta G_N$ is < 0 or > 1 . Since $\Delta G_N = \Delta G_\alpha + \Delta G_t$, this implies that ΔG_α and ΔG_t have opposite signs. Our model reproduces the clearly negative Φ -values for the mutations D23A in the CI2 helix and D38A in the protein L helix. Both mutations stabilize the helix (i.e. $\Delta G_\alpha < 0$), but destabilize tertiary interactions ($\Delta G_t > 0$).⁴

Conclusions

We have shown how to obtain a structural interpretation of Φ -values for multiple mutations within a single helix. Combined with any scale of helical propensities, our model shows how linear fitting of experimental data leads to two structural quantities: the extent of helix formation in the transition state, and the extent to which the helix interactions with neighboring tertiary structure are formed in the transition state. The method gives a simple physical interpretation of nonclassical Φ -values – nonclassical values arise if a mutation stabilizes a helix while destabilizing its interactions with neighboring parts of the protein, or vice versa. The model also explains how two different mutations at the same site can have different effects on the kinetics – this difference is traced back to different effects of the mutations on the intrinsic helix stability versus tertiary stability. Hence, this model appears to give simple physically consistent structural explanations for experimentally measured Φ -values.

⁴ In our previous article [53], we had erroneously stated that nonclassical Φ -values can arise for mutations that only shift the free energy of the denatured state, but not the free energy of the transition state and native state. This is not the case. Indeed, the Φ -value for these hypothetical mutations is 1 since $\Delta G_T = \Delta G_N$.

References

- [1] Itzhaki, L. S., Otzen, D. E., & Fersht, A. R. (1995). The structure of transition state for folding of chymotrypsin inhibitor 2 analysed by protein engineering methods: Evidence for a nucleation-condensation mechanism for protein folding. *J. Mol. Biol.* **254**, 260-288.
- [2] Villegas, V., Martinez, J. C., Aviles, F. X., & Serrano, L. (1998). Structure of the transition state in the folding process of human procarboxypeptidase A2 activation domain. *J. Mol. Biol.* **283**, 1027-1036.
- [3] Chiti, F., Taddei, N., White, P. M., Bucciantini, M., Magherini, F., Stefani, M., & Dobson, C. M. (1999). Mutational analysis of acylphosphatase suggests the importance of topology and contact order in protein folding. *Nat. Struct. Biol.* **6**, 1005-1009.
- [4] Ternström, T., Mayor, U., Akke, M., & Oliveberg, M. (1999). From snapshot to movie: Φ analysis of protein folding transition states taken one step further. *Proc. Natl. Acad. Sci. USA* **96**, 14854-14859.
- [5] Fulton, K. F., Main, E. R. G., Daggett, V., & Jackson, S. E. (1999). Mapping the interactions present in the transition state for unfolding/folding of FKBP12. *J. Mol. Biol.* **291**, 445-461.
- [6] Kim, D. E., Fisher, C., & Baker, D. (2000). A breakdown of symmetry in the folding transition state of protein L. *J. Mol. Biol.* **298**, 971-984.
- [7] McCallister, E. L., Alm, E., & Baker, D. (2000). Critical role of β -hairpin formation in protein G folding. *Nat. Struct. Biol.* **7**, 669-673.
- [8] Otzen, D. E., & Oliveberg, M. (2002). Conformational plasticity in folding of the split β - α - β protein S6: Evidence for burst-phase disruption of the native state. *J. Mol. Biol.* **317**, 613-627.
- [9] Hedberg, L., & Oliveberg, M. (2004). Scattered Hammond plots reveal second level of site-specific information in protein folding: $\Phi'(\beta^\ddagger)$. *Proc. Natl. Acad. Sci. USA* **101**, 7606-7611.
- [10] Anil, B., Sato, S., Cho, J. H., & Raleigh, D. P. (2005). Fine structure analysis of a protein folding transition state; distinguishing between hydrophobic stabilization and specific packing. *J. Mol. Biol.* **354**, 693-705.
- [11] Went, H. M., & Jackson, S. E. (2005). Ubiquitin folds through a highly polarized transition state *Protein Engineering, Design & Selection*, **18**, 229-237.
- [12] Wilson, C. J., & Wittung-Stafshede, P. (2005). Snapshots of a dynamic folding nucleus in zinc-substituted *Pseudomonas aeruginosa* azurin. *Biochemistry* **44**, 10054-10062.

- [13] Kragelund, B. B., Osmark, P., Neergaard, T. B., Schiodt, J., Kristiansen, K., Knudsen, J., & Poulsen, F. M. (1999) The formation of a native-like structure containing eight conserved hydrophobic residues is rate limiting in two-state protein folding of ACBP. *Nature Struct. Biol.* **9**, 594-601.
- [14] Gianni, S., Guydosh, N. R., Khan, F., Caldas, T. D., Mayor, U., White, G. W. N. DeMarco, M. L., Daggett, V., & Fersht, A. R. (2003). Unifying features in protein-folding mechanisms. *Proc. Natl. Acad. Sci. USA* **100**, 13286-13291.
- [15] Sato, S., Religa, T. L., Daggett, V., & Fersht, A. R. (2004). Testing protein-folding simulations by experiment: B domain of protein A. *Proc. Natl. Acad. Sci. USA* **101**, 6952-6956.
- [16] Teilum, K., Thormann, T., Caterer, N. R., Poulsen, H. I., Jensen, P. H., Knudsen, J., Kragelund, B. B., & Poulsen, F. M. (2005). Different secondary structure elements as scaffolds for protein folding transition states of two homologous four-helix bundles. *Proteins* **59**, 80-90.
- [17] Martinez, J. C., & Serrano, L. (1999). The folding transition state between SH3 domains is conformationally restricted and evolutionarily conserved. *Nature Struct. Biol.* **6**, 1010-1016.
- [18] Riddle, D. S., Grantcharova, V. P., Santiago, J. V., Alm, E., Ruczinski, I., & Baker, D. (1999). Experiment and theory highlight role of native state topology in SH3 folding. *Nature Struct. Biol.* **6**, 1016-1024.
- [19] Hamill, S. J., Steward, A., & Clarke, J. (2000). The folding of an immunoglobulin-like greek key protein is defined by a common-core nucleus and regions constrained by topology. *J. Mol. Biol.* **297**, 165-178.
- [20] Fowler, S.B., & Clarke, J. (2001). Mapping the folding pathway of an Immunoglobulin domain: Structural detail from Φ value analysis and movement of the transition state. *Structure* **9**, 355-366.
- [21] Cota, E., Steward, A., Fowler, S. B., & Clarke J. (2001). The folding nucleus of a fibronectin type III domain is composed of core residues of the immunoglobulin fold. *J. Mol. Biol.* **305**, 1185-1194.
- [22] Jäger, M., Nguyen, H., Crane, J.C., Kelly, J.W., & Gruebele, M. (2001). The folding mechanism of a β -sheet: The WW domain. *J. Mol. Biol.* **311**, 373-393.
- [23] Northey, J. G. B., Di Nardo, A. A., & Davidson, A. R. (2002). Hydrophobic core packing in the SH3 domain folding transition state. *Nature Struct. Biol.* **9**, 126-130.
- [24] Garcia-Mira, M. M., Böhringer, D., & Schmid, F. X. (2004). The folding transition state of the cold shock protein is strongly polarized. *J. Mol. Biol.* **339**, 555-569.

- [25] Matouschek, A., Kellis, J. T., Serrano, L., & Fersht, A. R. (1989). Mapping the transition state and pathway of protein folding by protein engineering. *Nature* **340**, 122-126.
- [26] Fersht, A. R. *Structure and mechanism in protein science* (W. H. Freeman, New York, 1999)
- [27] Ozkan, S. B., Bahar, I., & Dill, K. A. (2001). Transition states and the meaning of Φ -values in protein folding kinetics. *Nature Struct. Biol.* **8**, 765-769
- [28] Zarrine-Afsar, A., & Davidson, A. R. (2004). The analysis of protein folding kinetic data produced in protein engineering experiments. *Methods* **34**, 41-50.
- [29] Chang, I., Cieplak, M., Banavar, J. R., & Maritan, A. (2004). What can one learn from experiments about the elusive transition state? *Protein Sci.* **13**, 2446-2457.
- [30] Raleigh, D. P., & Plaxco, K. W. (2005). The protein folding transition state: what are Φ -values really telling us? *Protein and Peptide Letters* **12**, 117-122.
- [31] Li, A., & Daggett, V. (1994). Characterization of the transition state of protein unfolding by use of molecular dynamics: Chymotrypsin inhibitor 2. *Proc. Natl. Acad. Sci. USA* **91**, 10430-10434.
- [32] Li A, Daggett V. (1996). Identification and characterization of the unfolding transition state of chymotrypsin inhibitor 2 by molecular dynamics simulations. *J. Mol. Biol.* **257**, 412-429.
- [33] Lazaridis, T., & Karplus, M. (1997). "New View" of protein folding reconciled with the old through multiple unfolding simulations. *Science* **278**, 1928-1931.
- [34] Vendruscolo, M., Paci, E., Dobson, C. M., & Karplus, M. (2001). Three key residues form a critical contact network in a protein folding transition state. *Nature* **409**, 641-645.
- [35] Li, L., Shakhnovich, E. I. (2001). Constructing, verifying, and dissecting the folding transition state of chymotrypsin inhibitor 2 with all-atom simulations. *Proc. Natl. Acad. Sci. USA* **98**, 13014-13018.
- [36] Gsponer, J., & Caffisch, A. (2002). Molecular dynamics simulations of protein folding from the transition state. *Proc. Natl. Acad. Sci. USA* **99**, 6719-6724.
- [37] Paci, E., Vendruscolo, M., Dobson, C. M., & Karplus, M. (2002). Determination of a transition state at atomic resolution from protein engineering data. *J. Mol. Biol.* **324**, 151-163.
- [38] Guo, W., Lampoudi, S., & Shea, J.-E. (2003). Posttransition state desolvation of the hydrophobic core of the src-SH3 protein domain. *Biophys. J.* **85**, 61-69.

- [39] Settanni, G., Gsponer, J., & Caflisch, A. (2004). Formation of the folding nucleus of an SH3 domain investigated by loosely coupled molecular dynamics simulations. *Biophys. J.* **86**, 1691-1701.
- [40] Paci, E., Lindorff-Larsen, K., Dobson, C. M., Karplus, M., & Vendruscolo, M. (2005). Transition state contact orders correlate with protein folding rates. *J. Mol. Biol.* **352**, 495-500.
- [41] Salvatella, X., Dobson, C. M., Fersht, A. R., & Vendruscolo, M. (2005). Determination of the folding transition states of barnase by using Φ_I -value-restrained simulations validated by double mutant Φ_{IJ} -values. *Proc. Natl. Acad. Sci. USA* **102**, 12389-12394.
- [42] Chong, L. T., Snow, C. D., Rhee, Y. M., & Pande, V. S. (2005). Dimerization of the p53 oligomerization domain: Identification of a folding nucleus by molecular dynamics simulations. *J. Mol. Biol.* **345**, 869-878.
- [43] Hubner, I. A., Edmonds, K. A., & Shakhnovich, E. I. (2005). Nucleation and the transition state of the SH3 domain. *J. Mol. Biol.* **349**, 424-434.
- [44] Duan, J.X., & Nilsson, L. (2005). Thermal unfolding simulations of a multimeric protein – transition state and unfolding pathways. *Proteins* **59**, 170-182.
- [45] Daggett, V., Li, A., Itzhaki, L. S., Otzen, D. E., & Fersht, A. R. (1996). Structure of the transition state for folding of a protein derived from experiment and simulation. *J. Mol. Biol.* **257**, 430-440.
- [46] Day, R., & Daggett, V. (2005). Sensitivity of the folding/unfolding transition state ensemble of chymotrypsin inhibitor 2 to changes in temperature and solvent. *Protein Sci.* **14**, 1242-1252.
- [47] Settanni, G., Rao, F., & Caflisch, A. (2005). Φ -value analysis by molecular dynamics simulations of reversible folding. *Proc. Natl. Acad. Sci. USA* **102**, 628-633.
- [48] Lindorff-Larsen, K., Paci, E., Serrano, L., Dobson, C. M., & Vendruscolo, M. (2003). Calculation of mutational free energy changes in transition states for protein folding. *Biophys. J.* **85**, 1207-1214.
- [49] Goldenberg, D.P. Finding the right fold. *Nature Struct. Biol.* **6**, 987-990 (1999).
- [50] de los Rios, M. A., Daneshi, M., & Plaxco, K. W. (2005). Experimental investigation of the frequency and substitution dependence of negative Φ -values in two-state proteins. *Biochemistry* **44**, 12160-12167.
- [51] Schuler, B., Lipman, E., & Eaton, W. A. (2002). Probing the free-energy surface for protein folding with single molecule fluorescence spectroscopy. *Nature* **419**, 743-747.

- [52] Akmal, A., & Muñoz, V. (2004). The nature of the free energy barriers to two-state folding. *Proteins* **57**, 142-152.
- [53] Merlo, C., Dill, K. A., & Weikl, T. R. (2005). Φ values in protein-folding kinetics have energetic and structural components. *Proc. Natl. Acad. Sci. USA* **102**, 10171-10175.
- [54] Muñoz, V., & Serrano, L. (1994). Elucidating the folding problem of α -helical peptides using empirical parameters, II. Helix macrodipole effects and rational modification of the helical content of natural peptides. *J. Mol. Biol.* **245**, 275-296.
- [55] Muñoz, V., & Serrano, L. (1994). Elucidating the folding problem of α -helical peptides using empirical parameters III: Temperature and pH dependence. *J. Mol. Biol.* **245**, 297-308.
- [56] Lacroix, E., Viguera, A. R., & Serrano, L. (1998). Elucidating the folding problem of α -helices: Local motifs, long-range electrostatics, ionic strength dependence and prediction of NMR parameters. *J. Mol. Biol.* **284**, 173-191.
- [57] Pace, C. N., & Scholtz, J. M. (1998). A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75**, 422-427.
- [58] Serrano, L., Matouschek, A. & Fersht, A.R. (1992). The folding of an enzyme: III. Structure of the transition state for unfolding of barnase analysed by a protein engineering procedure *J. Mol. Biol.* **224**, 805-818.
- [59] Jemth, P., Day, R., Gianni, S., Khan, F., Allen, M., Daggett, V., & Fersht, A. R. (2005). The structure of the major transition state for folding of an FF domain from experiment and simulation. *J. Mol. Biol.* **350**, 363-378.
- [60] Zhou, Z., Huang, Y. Z., & Bai, Y. W. (2005). An on-pathway hidden intermediate and the early rate-limiting transition state of Rd-apocytochrome b(562) characterized by protein engineering. *J. Mol. Biol.* **352**, 757-764.
- [61] Scott, K. A., Randles, L. G., & Clarke, J. (2004). The folding of spectrin domains II: Φ -value analysis of R16. *J. Mol. Biol.* **344**, 207-221.
- [62] Sánchez, I. E., & Kiefhaber, T. (2003). Origin of unusual Φ -values in protein folding: evidence against specific nucleation sites. *J. Mol. Biol.* **334**, 1077-1085.
- [63] Fersht, A. R., & Sato, S. (2004). Φ -value analysis and the nature of protein-folding transition states. *Proc. Natl. Acad. Sci. USA* **91**, 10422-10425.
- [64] De los Rios, M., Muralidhara, B. K., Wildes, D., Sosnick, T. R., Marqusee, S., Wittung-Stafshede, P., Plaxco, K. W., & Ruczinski, I. (2006). On the precision of experimentally determined protein folding rates and Φ -values. *Protein Sci.* **15**, 553-563.

Table 1: Helix of the protein CI2

| mutation | Φ | ΔG_N | $\Delta G_\alpha^{\text{AGADIR}}$ | $\Delta G_\alpha^{\text{prop}}$ |
|----------|--------|--------------|-----------------------------------|---------------------------------|
| S12G | 0.29 | 0.8 | 0.28 | – |
| S12A | 0.43 | 0.89 | 0.14 | – |
| E15D | 0.22 | 0.74 | 0.13 | 0.29 |
| E15N | 0.53 | 1.07 | 0.57 | 0.25 |
| A16G | 1.06 | 1.09 | 0.82 | 1.0 |
| K17G | 0.38 | 2.32 | 0.80 | 0.74 |
| K18G | 0.7 | 0.99 | 0.75 | 0.74 |
| I20V | 0.4 | 1.3 | 0.14 | 0.2 |
| L21A | 0.25 | 1.33 | -0.01 | -0.21 |
| L21G | 0.35 | 1.38 | 0.26 | 0.79 |
| D23A | -0.25 | 0.96 | -0.41 | – |
| K24G | 0.1 | 3.19 | 0.12 | – |

Experimental Φ -values and stability changes ΔG_N are from Itzhaki et al.[1]. The change in intrinsic helix stability $\Delta G_\alpha^{\text{AGADIR}}$ is calculated with AGADIR [54,55,56], see Merlo et al. [53]. The change in intrinsic helix stability $\Delta G_\alpha^{\text{prop}}$ is calculated from the helix propensity scale of Pace and Scholtz [57]. The helix propensities of the residues are (in kcal/mol): Ala (A) 0, Leu (L) 0.21, Arg (R) 0.21, Met (M) 0.24, Lys (K) 0.26, Gln (Q) 0.39, Glu (E) 0.40, Ile (I) 0.41, Trp (W) 0.49, Ser (S) 0.50, Tyr (Y) 0.53, Phe (F) 0.54, Val (V) 0.61, His (H) 0.61, Asn (N) 0.65, Thr (T) 0.66, Cys (C) 0.68, Asp (D) 0.69, and Gly (G) 1. For the terminal residues 12, 13, 23, and 24 of the helix, the propensity scale is not applicable. We only consider mutations with $\Delta G_N > 0.7$ kcal/mol.

Table 2: Helix 2 of protein A

| mutation | Φ | ΔG_N | ΔG_α | tertiary contacts |
|----------|--------|--------------|-------------------|-------------------|
| A27G | 1.0 | 1.0 | 1.0 | – |
| A28G | 0.6 | 2.2 | 1.0 | Helix 1 |
| A29G | 1.1 | 1.0 | 1.0 | – |
| F31A | 0.3 | 3.9 | -0.54 | Helices 1, 3 |
| F31G | 0.5 | 4.7 | 0.46 | Helices 1, 3 |
| I32V | 0.6 | 1.2 | 0.2 | Helix 1 |
| I32A | 0.5 | 1.9 | -0.41 | Helix 1 |
| I32G | 0.6 | 3.4 | 0.59 | Helix 1 |
| A33G | 1.1 | 0.9 | 1.0 | – |
| A34G | 0.7 | 1.2 | 1.0 | – |
| L35A | 0.4 | 2.4 | -0.21 | Helices 1, 3 |
| L35G | 0.5 | 4.1 | 0.79 | Helices 1, 3 |

Experimental Φ -values and stability changes ΔG_N are from Sato et al. [15]. The change in intrinsic helix stability ΔG_α is calculated from the helix propensity scale of Pace and Scholtz [57]. The information whether tertiary contacts with helix 1 and 3 are affected by the mutations is taken from the contact matrix of protein A shown in Fig. 2. We only consider Φ -values for single-residue mutations with the wildtype sequence as reference state at those sites where multiple mutations have been performed. For example, we consider the Φ -values for the mutations I32V, I32A, and I32G in helix 2 of protein A, but not the Φ -values for V32A and A32G also given by Sato et al. [15]. However, we include the Φ -values for the Ala-Gly scanning mutants at the residue positions 27, 28, 29, 33, and 34 given in Table 1 of Sato et al. [15].

Table 3: Helix 3 of protein A

| mutation | Φ | ΔG_N | ΔG_α | tertiary contacts |
|----------|--------|--------------|-------------------|-------------------|
| A44G | -0.1 | 1.3 | 1.0 | – |
| L45A | 0.6 | 1.5 | -0.21 | Helix 2 |
| L45G | 0.3 | 4.4 | 0.79 | Helix 2 |
| L46A | 0.2 | 1.9 | -0.21 | Helix 1 |
| L46G | 0.3 | 4.0 | 0.79 | Helix 1 |
| A47G | 0.2 | 1.5 | 1.0 | – |
| A48G | 0.0 | 1.8 | 1.0 | Helix 2 |
| A49G | 0.2 | 3.6 | 1.0 | Helix 2 |
| A51G | 0.1 | 1.2 | 1.0 | – |
| L52A | 0.3 | 1.3 | -0.21 | Helix 2 |
| L52G | 0.1 | 3.8 | 0.79 | Helix 2 |
| A54G | 0.0 | 1.4 | 1.0 | – |

Experimental Φ -values and stability changes ΔG_N are from Sato et al. [15]. The change in intrinsic helix stability ΔG_α is calculated from helix propensities [57]. The information on tertiary contacts is taken from Fig. 2

Table 4: Helix of protein L

| mutation | Φ | ΔG_N | ΔG_α | tertiary contacts |
|----------|--------|--------------|-------------------|-------------------|
| A29G | 0.23 | 2.41 | 1.0 | Hairpin 1 |
| T30A | 0.08 | 1.31 | -0.66 | Hairpin 1 |
| S31G | 0.11 | 0.81 | 0.5 | — |
| E32G | 0.11 | 1.08 | 0.6 | Hairpin 1 |
| E32I | 0.05 | 1.25 | 0.01 | Hairpin 1 |
| A33G | 0.25 | 2.85 | 1.0 | Hairpin 1, 2 |
| Y34A | 0.05 | 2.57 | -0.53 | Hairpin 2 |
| A35G | 0.28 | 1.2 | 1.0 | — |
| Y36A | 0.27 | 2.54 | -0.53 | Hairpin 1 |
| A37G | 0.11 | 3.14 | 1.0 | Hairpin 2 |
| D38A | -0.39 | 0.98 | -0.69 | Hairpin 2 |
| D38G | -0.05 | 1.89 | 0.31 | Hairpin 2 |

Experimental Φ -values and stability changes ΔG_N are from Kim et al. [6]. The change in intrinsic helix stability ΔG_α is calculated from helix propensities [57]. The two β -hairpins of protein L are defined in the caption of Fig. 3. The information on tertiary interactions of helical residues with the hairpins is taken from this figure.

Table 5: Structural parameters, standard deviations, and correlation coefficients

| helix | tertiary contacts | χ_α | χ_t | SD | $ r $ |
|----------------------|-------------------|---------------|----------|------|---------------------|
| CI2 helix | all | 1.03 | 0.16 | 0.14 | 0.91 |
| helix 2 of protein A | all | 0.98 | 0.46 | 0.10 | 0.93 |
| | with helix 1 | 0.98 | 0.52 | 0.12 | 0.90 |
| helix 3 of protein A | all | -0.07 | 0.31 | 0.13 | 0.75 |
| | with helix 1 | -0.01 | 0.24 | 0.13 | 0.65 |
| | with helix 2 | -0.09 | 0.34 | 0.13 | 0.79 |
| helix of protein L | all | 0.30 | 0.06 | 0.15 | 0.63 |
| | with hairpin 1 | 0.21 | 0.15 | 0.10 | (0.30) ^a |
| | with hairpin 2 | 0.32 | -0.04 | 0.11 | 0.90 |

The structural parameters χ_α and χ_t , estimated standard deviations SD of the data points from the regression lines, and absolute values of the correlation coefficient r obtained in our model. The second column of the table indicates whether we consider all mutations for a helix, or only mutations affecting tertiary interactions with one structural element. The structural parameter χ_t then either indicates the overall degree of tertiary structure formation in the transition state, or the degree of tertiary structure formation with the given structural element. In both cases, we have included the ‘purely secondary’ mutations that do not affect tertiary interactions. The structural elements of protein A and L are defined in the Figs. 2 and 3. The standard deviation SD is estimated as $SD = \sqrt{(\sum_{i=1}^M d_i^2) / (M - 2)}$ where d_i is the vertical deviation of data point i from the regression line, and M is the number of data points. We estimate the errors in the structural parameters χ_α and χ_t , which result from experimental and modeling errors, as ± 0.05 for the CI2 helix and helix 2 of protein A, and as ± 0.1 for helix 3 of protein A and the protein L helix.

^aFor this data set, the correlation coefficient r is not a reasonable indicator of the modeling quality since the slope of the regression line is close to 0. The precise value of r is then dominated by the experimental errors in Φ . In our model, the slope of the regression line close to 0 indicates that the two structural parameters χ_α and χ_t have similar values, see eq. 3. The relatively small standard deviation SD of 0.10 for this data set shows that our model is in good agreement with the data.

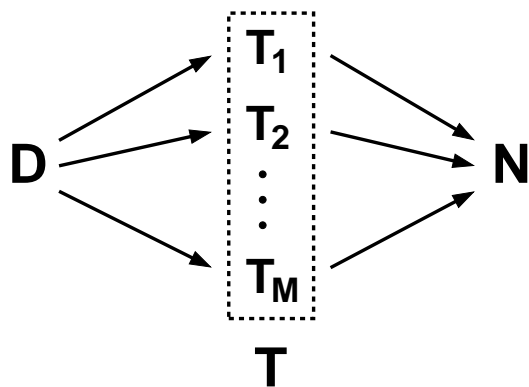


Fig. 1. In our model, the transition-state ensemble **T** consists of M transition-state conformations T_1, T_2, \dots, T_M . The arrows indicate the folding direction from the denatured state **D** to the native state **N** via the transition-state conformations.

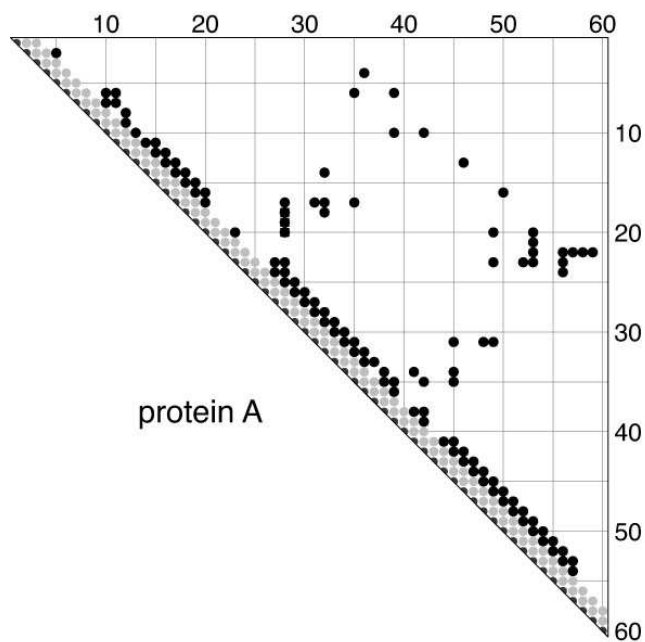


Fig. 2. Contact matrix of protein A. A black dot at position (i, j) of the matrix indicates that the two non-neighboring residues i and j are in contact in the native structure (protein data bank file 1SS1, model 1). Two residues here are defined to be in contact is the distance between any of their non-hydrogen atoms is smaller than the cutoff distance 4 Å. Protein A is an α -helical protein with three helices. Helix 1 consists of the residues 10 to 19, helix 2 of the residues 25 to 37, and helix 3 of the residues 42 to 56.

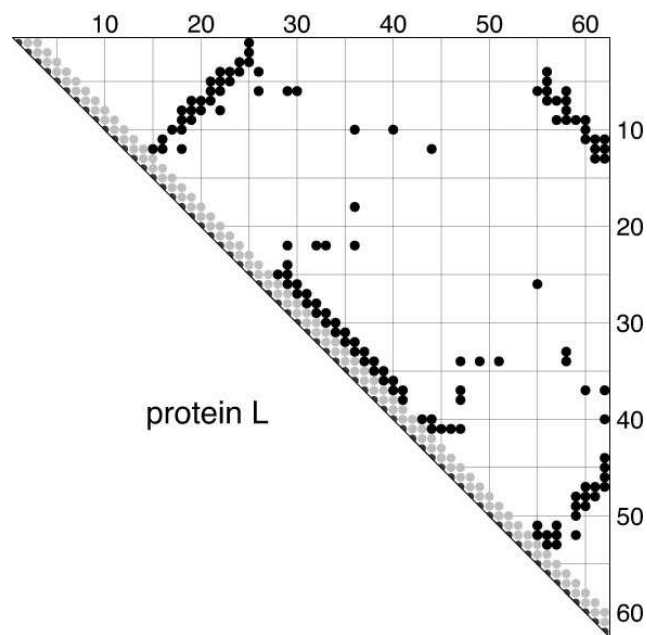


Fig. 3. Contact matrix of protein L (protein data bank file 1HZ6, residues A1 to A62). The structure of protein L consists of two β -hairpins at the termini, and an α -helix in between. The helix consists of the residues 26 to 40. The hairpin 1 at the N-terminus includes the residues 4 to 24, and the hairpin 2 at the C-terminus includes the residues 47 to 62.

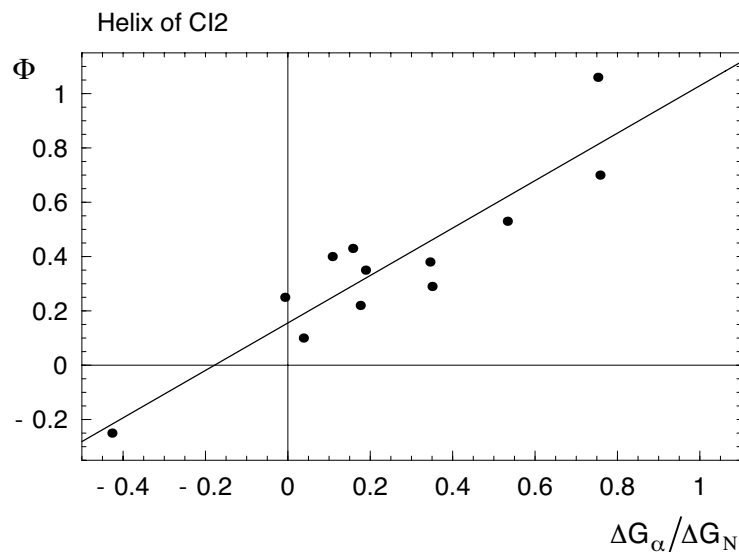


Fig. 4. Analysis of Φ -values for mutations in the helix of the protein CI2. The change in intrinsic helix stability ΔG_α for the 12 mutations has been calculated with AGADIR (see Table 1). We only consider mutations with experimentally measured stability changes $\Delta G_N > 0.7$ kcal/mol. The Pearson correlation coefficient of the 12 data points is 0.91. From the regression line $\Phi = 0.16 + 0.87\Delta G_\alpha/\Delta G_N$, we obtain the structural parameters $\chi_\alpha = 1.03 \pm 0.05$ and $\chi_t = 0.16 \pm 0.05$. The structural parameter χ_α close to 1 indicates that the helix is fully formed in the transition state. The parameter χ_t indicates that tertiary interactions are on average present in the transition state to a degree around 16 %

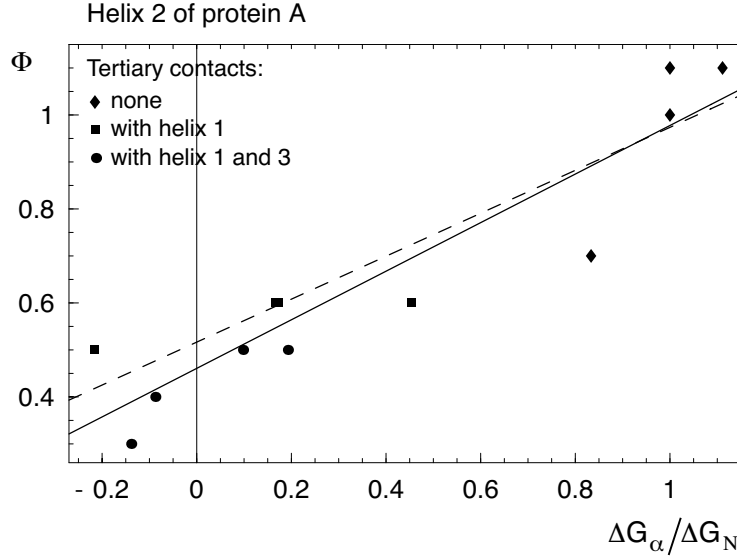


Fig. 5. Analysis of Φ -values for helix 2 of protein A. The solid line represents the regression line $\Phi = 0.46 + 0.52 \Delta G_{\alpha} / \Delta G_N$ for all points. The correlation coefficient of the data points is 0.93. The dashed line is the regression line $\Phi = 0.52 + 0.46 \Delta G_{\alpha} / \Delta G_N$ of the 8 data points for mutations of residues that have either no tertiary interactions or tertiary interactions with helix 1 (see also Table 2). The correlation coefficient of these data points is 0.90. From the regression lines and eq. (3), we obtain the structural parameters χ_{α} and χ_t shown in Table 5. The values of χ_{α} close to 1 indicate that the helix is fully formed in the transition state, and the values of χ_t close to 0.5 indicate that tertiary interactions are present to a degree of about 50 %.

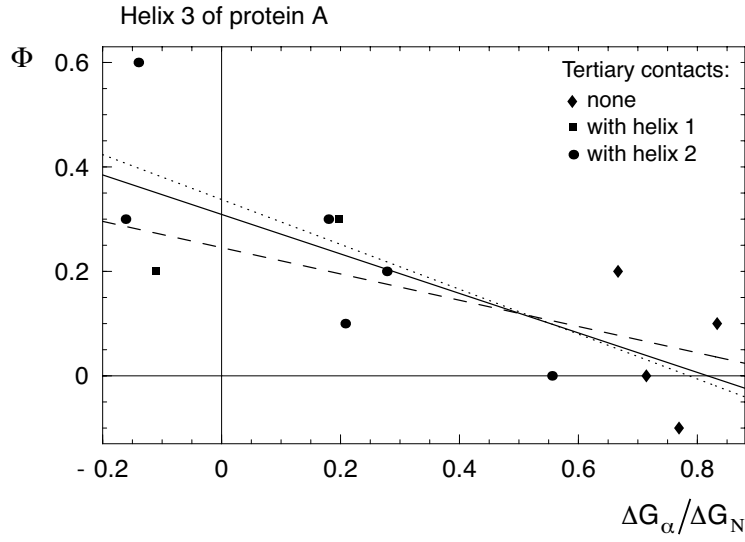


Fig. 6. Analysis of Φ -values for mutations in helix 3 of protein A. The solid line represents the regression line $\Phi = 0.31 - 0.38\Delta G_\alpha / \Delta G_N$ of all data points; the dashed line is the regression line $\Phi = 0.24 - 0.25\Delta G_\alpha / \Delta G_N$ of the data points for mutations that affect the tertiary interactions with helix 1 (or no tertiary interactions); and the dotted line is the regression line $\Phi = 0.34 - 0.43\Delta G_\alpha / \Delta G_N$ of data points for mutations that affect tertiary interactions interactions with helix 2 or no tertiary interactions). The absolute values of the correlation coefficient for these three data sets are $|r| = 0.75, 0.65$, and 0.79 , respectively (see Table 5).

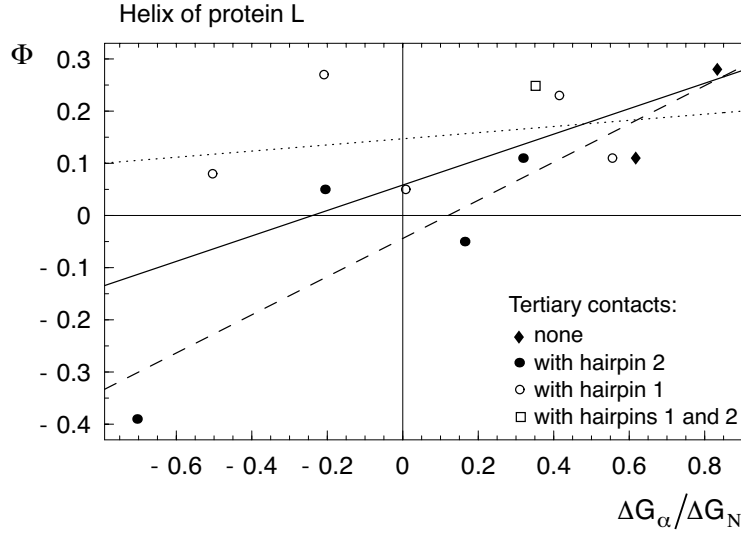


Fig. 7. Analysis of Φ -values for mutations in the helix of protein L. The solid line represents the regression line $\Phi = 0.06 + 0.24 \Delta G_{\alpha} / \Delta G_N$ for all data points; the dotted line is the regression line $\Phi = 0.15 + 0.06 \Delta G_{\alpha} / \Delta G_N$ of the 7 data points for mutations that affect tertiary interactions with hairpin 1 or none of the tertiary interactions (see also Table 4); and the dashed line is the regression line $\Phi = -0.04 + 0.37 \Delta G_{\alpha} / \Delta G_N$ of the 6 data points for mutations affecting tertiary interactions with hairpin 2 or none of the tertiary interactions.